Unleashing Text-to-Image Diffusion Models for Visual Perception

Year: 2023 | Citations: 288 | Authors: Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou

---

Abstract

Diffusion models (DMs) have become the new trend of generative models and have demonstrated a powerful ability of conditional synthesis. Among those, text-to-image diffusion models pre-trained on large-scale image-text pairs are highly controllable by customizable prompts. Unlike the unconditional generative models that focus on low-level attributes and details, text-to-image diffusion models contain more high-level knowledge thanks to the vision-language pre-training. In this paper, we propose VPD (Visual Perception with pre-trained Diffusion models), a new framework that exploits the semantic information of a pre-trained text-to-image diffusion model in visual perception tasks. Instead of using the pre-trained denoising autoencoder in a diffusion-based pipeline, we simply use it as a backbone and aim to study how to take full advantage of the learned knowledge. Specifically, we prompt the denoising decoder with proper textual inputs and refine the text features with an adapter, leading to a better alignment to the pre-trained stage and making the visual contents interact with the text prompts. We also propose to utilize the cross-attention maps between the visual features and the text features to provide explicit guidance. Compared with other pre-training methods, we show that vision-language pre-trained diffusion models can be faster adapted to downstream visual perception tasks using the proposed VPD. Extensive experiments on semantic segmentation, referring image segmentation, and depth estimation demonstrate the effectiveness of our method. Notably, VPD attains 0.254 RMSE on NYUv2 depth estimation and 73.3% oIoU on RefCOCO-val referring image segmentation, establishing new records on these two benchmarks. Code is available at https://github.com/wl-zhao/VPD.